

Improving Region Representation Learning from Urban Imagery with Noisy Long-Caption Supervision

Yimei Zhang^{1,2}, Guojiang Shen^{1,2}, Kaili Ning^{1,2}, Tongwei Ren³, Xuebo Qiu¹,
Mengmeng Wang^{1,2}, Xiangjie Kong^{1,2,*}

¹ College of Computer Science and Technology, Zhejiang University of Technology

² Zhejiang Key Laboratory of Visual Information Intelligent Processing

³ State Key Laboratory for Novel Software Technology, Nanjing University

{yimeizhang0229, gjshen1975, ningkaili, xueboqiu, wangmengmeng}@zjut.edu.cn;
rentw@nju.edu.cn; xjkong@ieee.org

Abstract

Region representation learning plays a pivotal role in urban computing by extracting meaningful features from unlabeled urban data. Analogous to how perceived facial age reflects an individual’s health, the visual appearance of a city serves as its “portrait”, encapsulating latent socio-economic and environmental characteristics. Recent studies have explored leveraging Large Language Models (LLMs) to incorporate textual knowledge into imagery-based urban region representation learning. However, two major challenges remain: i) difficulty in aligning fine-grained visual features with long captions, and ii) suboptimal knowledge incorporation due to noise in LLM-generated captions. To address these issues, we propose a novel pre-training framework called UrbanLN that improves Urban region representation learning through Long-text awareness and Noise suppression. Specifically, we introduce an information-preserved stretching interpolation strategy that aligns long captions with fine-grained visual semantics in complex urban scenes. To effectively mine knowledge from LLM-generated captions and filter out noise, we propose a dual-level optimization strategy. At the data level, a multi-model collaboration pipeline automatically generates diverse and reliable captions without human intervention. At the model level, we employ a momentum-based self-distillation mechanism to generate stable pseudo-targets, facilitating robust cross-modal learning under noisy conditions. Extensive experiments across four real-world cities and various downstream tasks demonstrate the superior performance of our UrbanLN.

Code — <https://github.com/YimeiZhang0229/UrbanLN>

Extended version — <https://arxiv.org/abs/2511.07062>

Introduction

Cities, as vital carriers for living, working, and entertainment, are becoming increasingly complex and diverse due to accelerated urbanization (Feng et al. 2024). Consequently, the integration of urban data and advanced deep learning methods for accurate inference of regional attributes has become a pivotal research focus in urban computing, empowering decision-makers with essential insights to guide strate-

gic actions in urban planning, sustainable development, and policy-making (Zhou et al. 2025; Liang et al. 2025; Zou et al. 2025).

Several research lines have emerged to integrate diverse spatial data for urban region representation. POIs have been extensively utilized as proxies for urban functions to assess city vitality, while dynamic population remains a fundamental variable in urban growth theories (Zhang, Long, and Cong 2022; Wu et al. 2022). However, recent research (Fan et al. 2023b) suggests that the visual appearance of the built environment may encapsulate richer information than functional attributes, residential density, or visitor activity patterns. As a result, researchers have increasingly focused on leveraging urban imagery for region representation learning (Xi et al. 2022; Chen et al. 2024b; Yong and Zhou 2024; Hao et al. 2025).

Previous studies primarily utilize segmentation model (Lee et al. 2021; Fan et al. 2023b) or visual model-based methods for urban imagery feature extraction (Huang et al. 2023). While effective, these approaches are typically limited in capturing high-level contextual semantics and latent urban knowledge embedded in complex visual scenes. In contrast, the remarkable performance of LLMs across various domains (Li et al. 2022; Feng et al. 2025), combined with their ability to process and generate language, has positioned them as powerful tools for multimodal learning. Recent advancements, such as UrbanCLIP (Yan et al. 2024) and UrbanVLP (Hao et al. 2025), leverage image captions generated by Multimodal LLMs (MLLMs), enabling deeper insights into urban imagery. Despite these efforts, two key challenges remain: 1) **Semantic Bottleneck in Long-Text Processing**: Captions generated by MLLMs embed extensive world knowledge and rich semantic information, providing valuable context for interpreting urban imagery. However, existing methods struggle to effectively process and distill fine-grained features from these often lengthy descriptions. 2) **Failure Knowledge Integration from Noisy Captions**: Captions generated by MLLMs often contain noise such as inaccuracies, omissions, or overly generic content (Li et al. 2025), which hinders effective knowledge integration during cross-modal pre-training. UrbanCLIP addresses this issue via manual calibration, which is time-consuming and not scalable. UrbanVLP introduces

*Corresponding author.

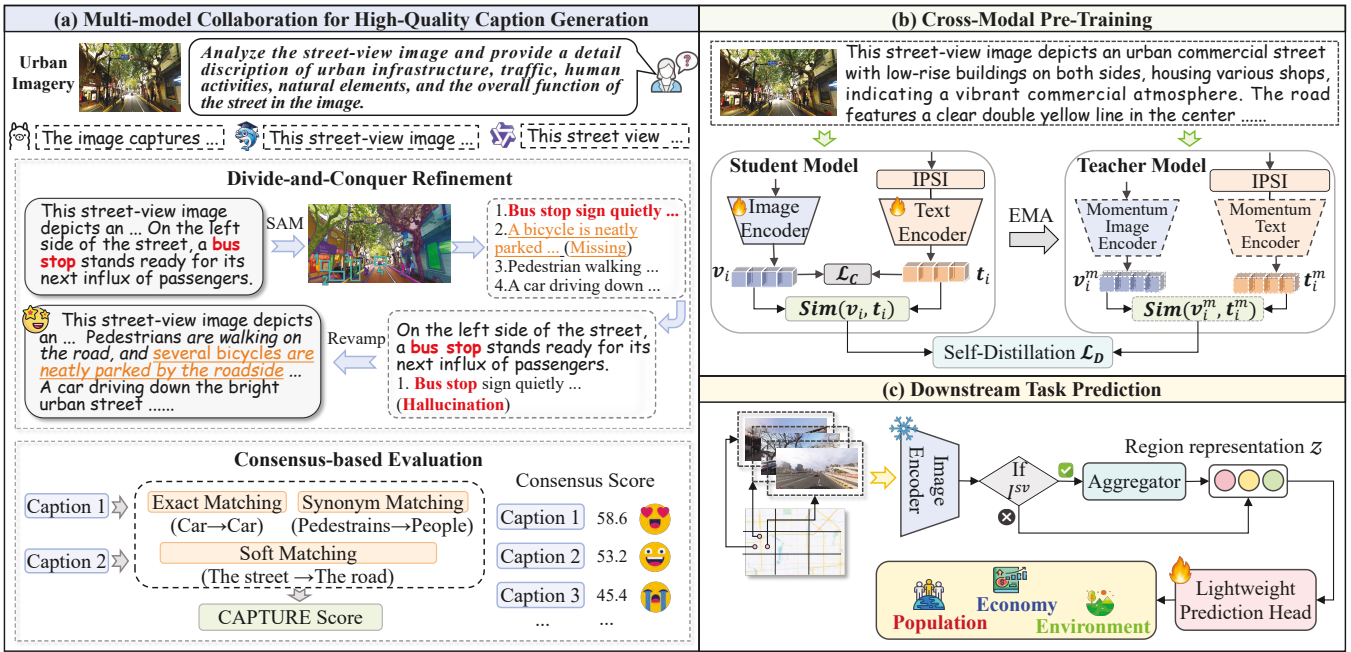


Figure 1: Overview of UrbanLN. The urban imagery input can be either satellite imagery or street-view imagery, with the methodology section primarily focusing on street-view imagery as a representative case.

scene segmentation ratios to guide caption generation; however, its reliance on fixed category schemas and overemphasis on quantitative cues leads to the loss of high-level semantic information. Furthermore, the use of rigid prompt templates limits linguistic diversity, resulting in homogeneous and low-information descriptions.

In this paper, we propose UrbanLN, a novel cross-modal pre-training framework that enhances urban region representation learning by modeling fine-grained semantics in long-text and mitigating caption noise for more effective knowledge integration. To address **Challenge 1**, we introduce an Information-Preserved Stretching Interpolation (IPSI) strategy, which enables effective alignment between long-form textual descriptions and fine-grained visual features in intricate urban imagery with minimal additional computational cost. To address **Challenge 2**, we propose a dual-level optimization scheme that mitigates noise from LLM-generated captions. At the data level, we integrate multi-MLLM captioning, a divide-and-conquer refinement operation, and consensus-based evaluation, which together promote semantic diversity and improve the reliability of captions. At the model level, we employ a momentum-based self-distillation mechanism, where a momentum-updated version of the student model serves as the teacher to generate pseudo-targets for additional supervision. These pseudo-targets can be seen as diverse “alternative views”, encouraging the student to learn robust cross-modal representations by differentiating informative cues from noisy content. In summary, this paper makes the following contributions:

- We introduce an information-preserved stretching interpolation strategy that relaxes input length constraints with minimal computational overhead. This enables the

model to extract fine-grained and holistic features from urban imagery, facilitating more comprehensive urban region representation.

- We propose a dual-level optimization strategy to mitigate noise introduced by LLM-generated captions. At the data level, we incorporate multi-MLLM captioning, divide-and-conquer refinement, and consensus-based evaluation to effectively filter out unreliable captions. At the model level, we employ a momentum-driven self-distillation approach to learn semantically robust representations under noisy supervision.
- Extensive experiments conducted across four real-world cities and diverse downstream tasks demonstrate that UrbanLN consistently outperforms state-of-the-art methods in urban region representation learning.

Preliminaries

Definition 1 (Urban Region) An urban region refers to a spatial unit obtained by partitioning a city based on a specific zoning strategy. Following previous studies (Yan et al. 2024; Hao et al. 2025), we obtain the region set \mathcal{R} by partitioning the urban area into a collection of equal-sized, non-overlapping square grids.

Definition 2 (Urban Imagery) Given a city, the urban imagery set is denoted by $\mathcal{I} = \mathcal{I}^{SV} \cup \mathcal{I}^{SI}$ with the street-view imagery set \mathcal{I}^{SV} and the satellite imagery set \mathcal{I}^{SI} . For each region $R_i \in \mathcal{R}$, it is associated with a set of u street-view images $I_i^{sv} = \{I_{i,1}^{sv}, I_{i,2}^{sv}, \dots, I_{i,u}^{sv}\} \subset \mathcal{I}^{SV}$ and a satellite image $I_i^{si} \in \mathcal{I}^{SI}$.

Problem Statement (Urban Region Representation Learning) Given the region set \mathcal{R} with its associated urban

imagery set \mathcal{I} , our goal is to learn the visual representation that serves as region representation, which can be leveraged for a variety of downstream tasks, such as socioeconomic indicator prediction.

Methodology

As illustrated in Figure 1, UrbanLN comprises three key components: 1) a multi-model collaboration caption generation pipeline to produce high-quality texts; 2) a cross-modal pre-training framework designed for efficient training of a powerful urban image encoder; and 3) a prediction module for various downstream tasks by simply fine-tuning a lightweight prediction head.

Multi-model Collaboration for High-Quality Caption Generation

To achieve a more comprehensive understanding of urban imagery, we first leverage MLLMs to generate semantically rich and context-aware descriptions for each image. These captions go beyond surface-level visual elements by incorporating the model’s world knowledge and reasoning capabilities, capturing key attributes that reflect the characteristics of urban environments. However, captions generated by MLLMs are often prone to hallucinations and information homogenization (Ye et al. 2025). To tackle this issue, we propose a multi-model collaboration approach aimed at generating diverse, comprehensive, and reliable captions by employing three key strategies as follows.

Multi-MLLM Captioning. We utilize multiple MLLMs, where each model independently serves as a captioning agent to generate multiple long captions. The diversity in different MLLMs not only mitigates potential semantic shifts introduced by individual models but also generates rich textual data across the entire dataset, which can be seen as a form of text data augmentation, similar to language rewriting (Fan et al. 2023a), thereby promoting subsequent cross-modal pre-training.

Divide-and-Conquer Refinement. To alleviate the issues of hallucinations and visual element omissions, we incorporate fine-grained visual details to reinforce the completeness and accuracy of the captions. Firstly, we employ the Segment Anything Model (SAM) (Kirillov et al. 2023) for image segmentation, and apply a maximal rectangle algorithm to reduce overlap among the remaining regions, resulting in a set of cropped bounding boxes that represent salient visual elements. Secondly, for each detected salient visual element, we utilize the respective MLLM to generate concise local short captions that complement the long caption by addressing potentially omitted visual details. Thirdly, to further suppress hallucinations, we introduce a phrase-level filtering mechanism. Visual element phrases are first extracted from both original long and local short captions using the Factual parser (Li et al. 2023), which is a state-of-the-art model for text scene graph parsing. These phrases are then scored by OWLv2 (Minderer, Gritsenko, and Hounsby 2023), an open-vocabulary object detection model. Phrases with a score below 0.01 are discarded, reducing the inclusion of hallucinated content.

Finally, we reuse the same MLLM that reasonably absorbs detailed information from local short captions based on image content and original long caption structures to generate more complete and natural image descriptions.

Consensus-based Evaluation. Following the above refinement process, the quality of the caption is improved. However, in the absence of ground-truth captions, existing quantitative metrics fall short in accurately assessing their quality. Inspired by real-world decision-making processes, where consensus among multiple capable and independent individuals is often considered reliable, we propose a consensus-based evaluation method. This method uses inter-model consistency as an indicator of caption reliability under unsupervised settings.

Specifically, we adopt the CAPTURE metric (Dong et al. 2024) to measure the similarity between two candidate captions, as it leverages structured visual semantics extracted by the Factual parser, enabling fine-grained alignment of objects, attributes, and relations, rather than relying on generic sentence-level embeddings. Following its procedure, we extract core information such as objects, attributes, and relations from the descriptions using the Factual parser, and remove redundant or irrelevant content. Then, for the extracted core information, using three matching strategies: 1) exact matching (checking if the core elements in the two captions match); 2) synonym matching (checking if two candidate captions’ synonym sets overlap); and 3) soft matching (encoding and computing the similarity for remaining unmatched phrases). Next, calculating the precision and recall rates for each type of core information, where the soft matching score is used as a complement to exact matched and synonym matched relationships. More details can be found in (Dong et al. 2024). Finally, the CAPTURE metric is obtained as follows:

$$\text{CAPTURE} = \frac{\alpha F1_{obj} + \beta F1_{attr} + \gamma F1_{rel}}{\alpha + \beta + \gamma}, \quad (1)$$

where α , β and γ are scale factors, and $F1_*$ stands for the $F1$ score of each type of core information.

The consensus score of each caption is then computed by averaging its CAPTURE scores with all other captions generated for the same image. A higher consensus score indicates stronger alignment with the shared understanding across multiple MLLMs, reflecting greater inter-model agreement. Finally, we select the caption candidate with the highest consistency score for each image to obtain a high-quality urban image-text pair dataset, denoted as $\mathcal{D} = \{\mathcal{I}, \mathcal{T}\}$.

Cross-Modal Pre-Training

To effectively transfer the knowledge embedded in the long captions into the visual representations, we adopt the Contrastive Language-Image Pre-training (CLIP) model (Radford et al. 2021) as the backbone to align text and image.

Information-Preserved Stretching Interpolation. Although CLIP has demonstrated strong capabilities in cross-modal representation learning, the token limitation of its text

encoder (77 tokens) drastically reduces its performance in processing long text inputs (Wu et al. 2024a; Zheng et al. 2024; Zhang et al. 2024). This limitation severely constrains its application in tasks requiring complex textual descriptions, especially in fine-grained urban region representation learning, where the average length of each caption exceeds 100 words. Consequently, directly applying CLIP to our task is not impractical.

To overcome CLIP’s limitations in handling long texts, we propose an Information-Preserved Stretching Interpolation (IPSI) strategy. This approach is inspired by prior studies (Zhang et al. 2024), which demonstrate that the lower positional embeddings (i.e., the first 20) in CLIP are well-trained and effectively capture absolute positional information. Directly interpolating these embeddings may disrupt the established positional representation, resulting in degraded performance and higher computational costs. To balance stability and computational efficiency, we preserve the first 20 positions and apply interpolation only to the remaining 57 positions. The detailed process is as follows:

$$P^*(e) = \begin{cases} P(e), & e \leq 20; \\ (1 - \omega)P(\lfloor \frac{e}{\lambda} \rfloor) + \omega P(\lceil \frac{e}{\lambda} \rceil), & \text{otherwise,} \end{cases} \quad (2)$$

where λ is the interpolation ratio, and $\omega = \frac{e}{\lambda} - \lfloor \frac{e}{\lambda} \rfloor$, which is used as a weighting factor to ensure smooth transition of the new position.

Momentum-based Self-Distillation. Despite our efforts to improve caption quality, errors or hallucinations remain non-trivial issues. Furthermore, due to the high visual similarity among certain urban neighborhoods, some captions selected as negatives may still accurately describe the visual content of the corresponding urban images. To tackle these issues, we introduce a momentum-based self-distillation method designed to improve learning under noisy supervision. This method keeps a momentum version (He et al. 2020; Li et al. 2021) of the student model as the teacher. During training, the teacher’s parameters are updated at each iteration using the Exponential Moving Average (EMA) of the student’s, and the student minimizes a contrastive loss \mathcal{L}_C , while aligning its outputs with those of the teacher via a distillation loss \mathcal{L}_D .

Specifically, we maintain two dynamic queues $\{\mathbf{v}_k^m, \mathbf{t}_k^m\}_{k \in K}$ that store the most recent K image-text representations encoded by the teacher model. Then, the similarity of the image-text and text-image can be formulated as:

$$p_k^{i2t}(I_i) = \frac{\exp(s_{\mathbf{v}_i, \mathbf{t}_k} / \tau)}{\sum_{k=1}^{K+N} \exp(s_{\mathbf{v}_i, \mathbf{t}_k} / \tau)}, \quad (3)$$

$$p_k^{i2t}(T_i) = \frac{\exp(s_{\mathbf{t}_i, \mathbf{v}_k} / \tau)}{\sum_{k=1}^{K+N} \exp(s_{\mathbf{t}_i, \mathbf{v}_k} / \tau)},$$

where N is the batch size, $s_{\mathbf{v}_i, \mathbf{t}_k} = \cos\langle \mathbf{v}_i, \mathbf{t}_k^m \rangle$, $s_{\mathbf{t}_i, \mathbf{v}_k} = \cos\langle \mathbf{t}_i, \mathbf{v}_k^m \rangle$, and τ is a learnable temperature parameter. The contrastive loss is:

$$\mathcal{L}_C = -\frac{1}{2N} \left(\sum_{i=1}^N \log p^{i2t}(I_i) + \sum_{i=1}^N \log p^{i2t}(T_i) \right). \quad (4)$$

Next, we incorporate a distillation mechanism that allows the student to learn from the teacher’s more stable pseudo-targets. We firstly compute the image-text similarity $s_{\mathbf{v}_i, \mathbf{t}_i}^m = \cos\langle \mathbf{v}_i^m, \mathbf{t}_i^m \rangle$ from the teacher, and obtain pseudo-targets q^{i2t} and q^{i2i} using the same formulation as Eq. (3) but replacing s in the original formulation with s^m . The distillation loss is then computed as:

$$\mathcal{L}_D = \frac{1}{2} \mathbb{E}_{(I, T) \sim \mathcal{D}} [\text{KL}(q^{i2t}(I) \| p^{i2t}(I)) + \text{KL}(q^{i2i}(T) \| p^{i2i}(T))], \quad (5)$$

where KL denotes the Kullback-Leibler divergence. The pseudo-targets generated by the momentum teacher offer additional perspectives beyond the original image-text pairs, guiding the student to learn representations that are robust to semantic-preserving transformations under noisy conditions. This encourages the model to capture shared semantic clues across diverse samples rather than overfitting noise. The final loss for pre-training is:

$$\mathcal{L} = (1 - \mu)\mathcal{L}_C + \mu\mathcal{L}_D, \quad (6)$$

where $\mu \in (0, 1)$ controls the balance between contrastive learning and teacher-guided supervision.

Downstream Task Prediction

Through optimizing the loss function in Eq. (6), we obtain a knowledge-enhanced image encoder. Given any urban imagery I_i^{sv} or I_i^{st} corresponding to a specific region R_i , we use the frozen image encoder to extract a compact feature representation \mathbf{z}_i , which serves as the urban region representation¹. For each downstream task, we fine-tune a lightweight prediction head (e.g., a multi-layer perceptron) to produce the output as $\hat{Y}_i = MLP(\mathbf{z}_i)$.

Experiment

Experimental Setup

Datasets. We collect satellite and street-view images for Beijing (BJ), Shanghai (SH), and Shenzhen (SZ) from Baidu Maps. The corresponding data for New York (NY) are sourced from (Yong and Zhou 2024). For downstream tasks, we evaluate our UrbanLN on the following urban indicators: population (**Pop**), gross domestic product (**GDP**), nighttime light intensity (**Night**), number of restaurant comments (**Com**), carbon emissions (**CO₂**), number of POIs (**POI**), and crime incidence (**Crime**). More details can be found in the Supplementary Material. All indicator data have been transformed into the logarithmic scale using the formula $Y = \ln(Y_{\text{original}} + 1)$. In the prediction phase, every downstream task dataset is randomly divided into training, validation, and test sets in a ratio of 6:2:2.

Implementation Details. We employ several MLLMs for long caption generation, including LLaMA-Adapter V2 (Gao et al. 2023), ShareGPT4V-7B (Chen et al. 2024a), Qwen2.5-VL-7B (Wang et al. 2024), DeepSeek-VL2-tiny

¹For multiple street-view images of the same region, we use average pooling of visual features to represent the region.

Task →	Pop			GDP			Night			Com			CO ₂		
	R^2 ↑	RMSE↓	MAE↓	R^2 ↑	RMSE↓	MAE↓	R^2 ↑	RMSE↓	MAE↓	R^2 ↑	RMSE↓	MAE↓	R^2 ↑	RMSE↓	MAE↓
ResNet-18	0.323	0.912	0.728	0.093	1.702	1.402	0.105	0.875	0.744	0.255	2.612	2.082	0.312	0.725	0.521
Tile2Vec	0.446	0.903	0.716	0.165	1.565	1.350	0.268	0.660	0.547	0.296	2.489	2.102	0.336	0.665	0.517
PG-SimCLR	0.549	0.802	0.612	0.316	1.450	1.163	0.334	0.669	0.538	0.429	2.337	1.862	0.467	0.642	0.452
RemoteCLIP	0.422	0.908	0.684	0.245	1.524	1.241	0.128	0.766	0.607	0.338	2.516	2.059	0.378	0.693	0.501
UrbanCLIP	0.595	0.761	0.556	0.358	1.410	1.121	0.426	0.621	0.503	0.504	2.226	1.764	0.480	0.619	0.492
UrbanVLP	<i>0.619</i>	<i>0.728</i>	<i>0.524</i>	<i>0.372</i>	<i>1.390</i>	<i>1.108</i>	<i>0.454</i>	<i>0.606</i>	<i>0.493</i>	<i>0.555</i>	<i>2.108</i>	<i>1.687</i>	<i>0.487</i>	<i>0.615</i>	<i>0.492</i>
UrbanLN+SV	0.705	0.686	0.491	0.440	1.373	1.043	<u>0.514</u>	<u>0.562</u>	<u>0.449</u>	<u>0.591</u>	1.976	<u>1.572</u>	<u>0.677</u>	<u>0.496</u>	<u>0.390</u>
Improvement	13.9%	5.8%	6.3%	18.3%	1.2%	5.9%	13.2%	7.3%	8.9%	6.5%	6.3%	6.8%	39.0%	19.3%	13.7%
UrbanLN+SI	<u>0.686</u>	<u>0.714</u>	<u>0.496</u>	<u>0.437</u>	<u>1.380</u>	<u>1.057</u>	0.519	0.553	0.447	0.597	2.005	1.568	0.688	0.478	0.383
Improvement	10.8%	1.9%	5.3%	17.5%	0.7%	4.6%	14.3%	8.7%	9.3%	7.6%	4.9%	7.1%	41.3%	22.3%	15.3%

Table 1: Downstream tasks prediction results on the BJ dataset. The abbreviations “+SV” and “+SI” denote the inclusion of street-view images and satellite images as urban imagery. The best, second-best, and third-best results are highlighted in bold, underlined, and italic, respectively. The “Improvement” row indicates the relative gain (%) over the best-performing baseline.

Model	Pop	Crime	POI
ResNet-18	-0.079	0.050	0.160
Tile2Vec	0.096	0.135	0.192
PG-SimCLR	0.375	0.196	–
RemoteCLIP	0.364	0.154	0.328
UrbanCLIP	0.448	0.251	0.425
MuseCL	0.521	<i>0.368</i>	<i>0.542</i>
UrbanVLP	<i>0.534</i>	<u>0.467</u>	<u>0.583</u>
UrbanLN+SV	<u>0.676</u>	0.723	0.650
Improvement	26.6%	54.8%	11.5%
UrbanLN+SI	0.725	0.323	0.521
Improvement	35.8%	-30.8%	-10.6%

Table 2: Prediction results on the NY dataset (R^2 metric).

(Wu et al. 2024b), and InternVL3-8B (Zhu et al. 2025). We set $\alpha, \beta, \gamma = 5, 5, 2$ following the default settings in (Dong et al. 2024). During pre-training, we use the CLIP model with ViT-B/16 backbone. The interpolation ratio λ is set to 4, extending the maximum input length to 248, and the weight μ is set to 0.5. The momentum parameter for updating the momentum teacher and the queue length are empirically set to 0.995 and 4096. To optimize UrbanLN, we utilize the AdamW optimizer and initialize the learning rate at $1e-7$.

Baselines. We compare UrbanLN with a variety of baselines commonly used in imagery-based urban region representation learning. These include ResNet-18 (He et al. 2016), Tile2Vec (Jean et al. 2019), PG-SimCLR (Xi et al. 2022), RemoteCLIP (Liu et al. 2024), UrbanCLIP (Yan et al. 2024), MuseCL (Yong and Zhou 2024) and UrbanVLP (Hao et al. 2025). MuseCL is restricted to the NY dataset owing to missing mobility data in BJ, SH, and SZ. The details of the baselines are in the Supplementary Material.

Evaluation Metrics. For downstream tasks evaluation, we employ widely used metrics including coefficient of determination (R^2), Root Mean Squared Error (RMSE), and Mean Absolute Error (MAE).

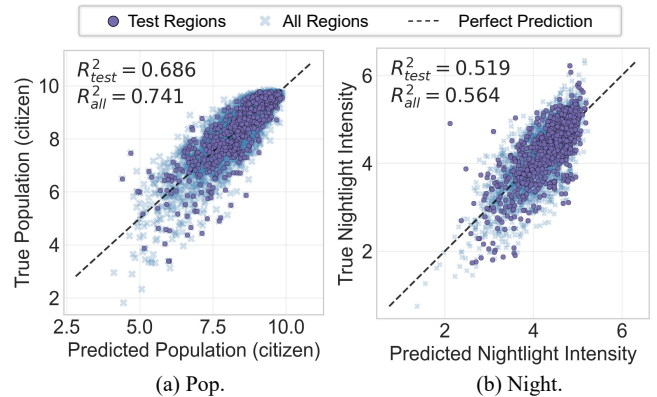
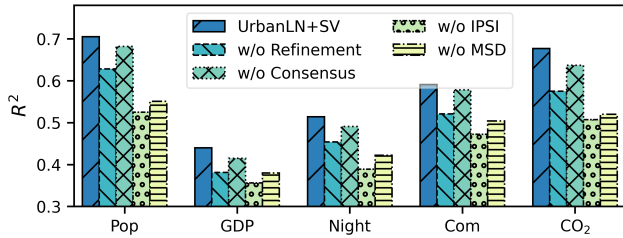


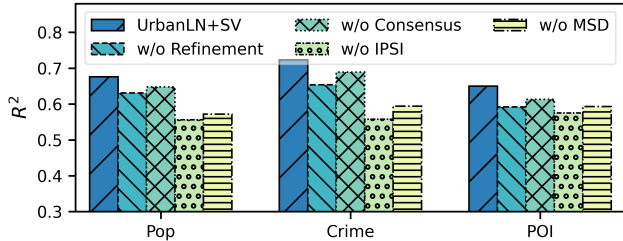
Figure 2: Prediction versus the ground truth on the BJ dataset using satellite imagery. The dotted line is at 45° . R^2_{test} and R^2_{all} correspond to the results of testing regions (purple dots) and all regions (blue crosses), respectively.

Comparative Results

The experimental results on the BJ dataset are shown in Table 1. Compared with the strongest baselines, the proposed UrbanLN achieves average improvements of 18.23%, 7.84%, and 8.32% in terms of R^2 , RMSE, and MAE, respectively. PG-SimCLR introduces POI semantic similarity and geographical proximity, thus yielding moderate improvements. RemoteCLIP, a foundation model for remote sensing images, excels in image-text retrieval and object counting but shows limited capability in urban understanding tasks. UrbanCLIP and UrbanVLP enhance image representations through textual description; however, their performance is constrained by limited semantic extraction, as well as hallucinations and homogenized expressions in the generated captions. By contrast, UrbanLN enables the input of long texts to capture fine-grained features in urban imagery, facilitating a more comprehensive understanding of urban environments. Furthermore, the proposed noise optimization strategy is applied at both the data and model levels, which enhances UrbanLN’s ability to integrate knowledge from



(a) Street-view imagery on BJ.



(b) Street-view imagery on NY.

Figure 3: Results of ablation study on R^2 metric.

LLM-generated captions, thereby contributing to its superior performance. To further assess the predictive power of UrbanLN, we compare the predicted and true urban indicators for all regions in the BJ dataset. As shown in Figure 2, UrbanLN closely replicates the true values for most regions.

We also include datasets from additional cities to evaluate the adaptability of our model. The R^2 results for NY are summarized in Table 2. Due to space limitations, results on additional evaluation metrics and the SH and SZ datasets are provided in the Supplementary Material. Our model consistently demonstrates superior performance across all downstream tasks in both SH and SZ. Notably, on the NY dataset, the UrbanLN+SV delivers the best results across all prediction tasks, with an average improvement of 30.97%. In contrast, UrbanLN+SI shows relatively suboptimal results in crime and POI prediction tasks. This performance gap may stem from the inherent limitations of satellite imagery and the lower spatial granularity of this dataset. By comparison, baselines such as MuseCL and UrbanVLP, which incorporate large-scale street-view images, achieve competitive performance, benefiting from their enhanced ability to model internal urban structures and human activity patterns.

Model Analysis

Ablation Study. To evaluate the effectiveness of the key components in UrbanLN, we conduct ablation studies across four critical variants. Results are presented in Figure 3, while additional results from the analysis of satellite imagery can be found in the Supplementary Material. These findings are summarized as follows: **1) w/o Refinement:** Removing the divide-and-conquer refinement module results in a 10.45% average performance drop, which highlights the importance of visual verification, local detail enhancement, and hallucination filtering in improving caption quality. **2) w/o Consensus:** Replacing the consensus-based caption selection with

Imagery	μ	City			
		BJ	SH	SZ	NY
Street-view	0.3	0.687	0.655	0.662	0.658
	0.4	0.697	0.666	0.678	0.669
	0.5	0.705	0.669	0.681	0.676
	0.6	0.692	0.655	0.678	0.674
Satellite	0.3	0.663	0.654	0.659	0.702
	0.4	0.679	0.660	0.673	0.710
	0.5	0.686	0.661	0.677	0.725
	0.6	0.682	0.665	0.678	0.721

Table 3: Impact of hyperparameter μ on population prediction across four cities on R^2 metric.

random choice leads to inferior performance, supporting our hypothesis that inter-model agreement serves as a reliable proxy for caption quality in the absence of ground-truth references. **3) w/o IPSI:** Omitting the IPSI strategy causes the most significant performance degradation across all scenarios, with an average reduction of 26.45%, marking its critical role in handling long-text inputs. By preserving rich information during knowledge transformation, this method ensures high semantic fidelity, which is particularly essential for complex urban scenarios. **4) w/o MSD:** Removing the momentum-based self-distillation module results in a notable performance drop, underscoring the importance of generating pseudo-targets in guiding the student model to learn semantically invariant representations under noisy conditions.

Hyperparameter Analysis. We explore the impact of pre-training loss weight μ on model performance, with results presented in Table 3. UrbanLN shows minimal performance variation across different values of μ , but achieves a slight improvement when $\mu = 0.5$. Therefore, we adopt $\mu = 0.5$ as the default setting.

Model Scale and Inference Efficiency. We compare the parameter scale and inference efficiency of UrbanLN with two strong baselines: UrbanCLIP and UrbanVLP. Since the caption generation pipeline is a part of data preprocessing, it is excluded from the analysis. As illustrated in Figure 4, all three models maintain comparable total parameter counts. However, UrbanLN substantially reduces the number of trainable parameters. This advantage can be attributed to the strategic integration of the pretrained CLIP model, which allows effective adaptation to urban region representations with minimal fine-tuning. For inference speed, we adopt Frames Per Second (FPS) as the evaluation metric, which is widely used in the computer vision area. While UrbanCLIP achieves the highest FPS, it does so at the expense of prediction accuracy. In contrast, UrbanLN strikes a more favorable balance between accuracy and efficiency, underscoring its potential for real-time urban applications.

Transferability Testing

To assess the capability of our model in the practical case, we pre-train the UrbanLN in a source city and evaluate its

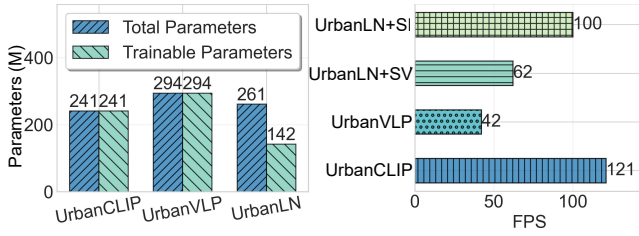
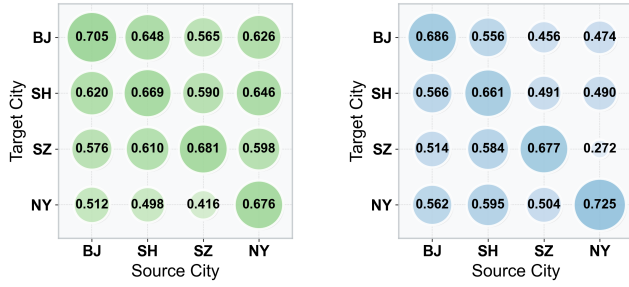


Figure 4: Comparison of parameters and inference speed.



(a) Street-view Imagery

(b) Satellite Imagery

Figure 5: The R^2 for the transferability test on street-view and satellite imagery-based population prediction.

transferability by applying it to other cities. The experimental results presented in Figure 5 indicate that, despite the inherent differences among cities, the model seems to capture fundamental commonalities across diverse urban environments, consistently achieving high prediction accuracy. This performance can be attributed to the effective alignment of visual and textual representations, allowing the model to extract universal functional semantics from urban images.

Latent Space Exploration

This case study aims to disentangle the information encoded in urban region representations and provide an intuitive illustration of the model’s capabilities in modeling city features. Using Shenzhen as an example, we first project the learned representations via principal component analysis and apply K-means clustering to partition the city into three distinct categories. Then, we perform a density-based analysis using population data to examine the characteristic patterns of each cluster.

As shown in Figure 6, the clusters exhibit clear spatial distinctions aligned with urban development patterns. Cluster 1 (green) shows a sharp, narrow peak in the KDE curves, indicating dense central urban areas with high population. Street-view images from this cluster reveal high-rise buildings, busy streets, and commercial infrastructure, supporting this interpretation. Cluster 0 (blue), which overlaps with Clusters 1 and 2, likely represents transitional zones between urban cores and outlying areas. Cluster 2 (orange) corresponds to suburban or rural regions with lower population. These findings demonstrate the model’s ability to capture meaningful socio-economic structures within urban environments.

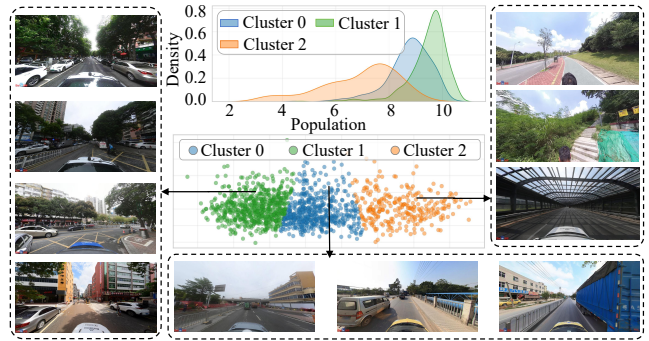


Figure 6: Case study on SZ using street-view imagery.

Related Work

Early studies primarily relied on supervised learning for specific tasks (Yeh et al. 2020). However, these “one-to-one” approaches not only require large amounts of labeled data but also suffer from limited transferability. To address this limitation, recent studies (Li et al. 2024; Zhang, Long, and Cong 2022; Sun et al. 2025) have shifted toward self-supervised representation learning for urban region analysis. This general-purpose approach eliminates the need to train specialized models for each downstream task, thereby enhancing efficiency and improving model generalizability (Liang et al. 2025).

Satellite imagery serves as a crucial visual data source for studying the spatial characteristics of urban regions and has been widely utilized in urban understanding (Burke et al. 2021). PG-SimCLR (Xi et al. 2022) and ReFound (Xiao et al. 2024) leveraged a combination of visual information from satellite imagery and semantic information from POIs to construct urban region representations. UrbanCLIP (Yan et al. 2024) introduced textual descriptions as a complementary modality to enhance the visual information contained in images. Moreover, street-view images have gained popularity in urban understanding due to their rich visual details. Vision-LSTM (Huang et al. 2023) used an LSTM-based network to integrate diverse street images into the satellite image domain, while UrbanVLP (Hao et al. 2025) integrated two visual modalities through token-level contrastive learning to capture multi-granularity information.

Conclusion

In this work, we propose UrbanLN, a cross-modal pre-training framework designed to improve urban region representation by addressing two key challenges in LLM-guided urban imagery analysis: the difficulty of extracting fine-grained semantics from long-form textual descriptions and the presence of noise in LLM-generated captions. To this end, UrbanLN introduces a long-text-aware strategy to preserve detailed semantic alignment and a hierarchical noise mitigation mechanism operating at both the data and model levels. The experiment results demonstrate that our model achieves state-of-the-art performance across various downstream tasks in four real-world datasets.

Ethical Statement

This study uses satellite imagery and street-view images obtained from the Baidu platform solely for academic research purposes. The spatial resolution of the satellite imagery is insufficient to identify individuals and is used solely to detect environmental changes associated with human activities. Street-view images were crawled and downloaded from publicly accessible content on the platform, with key identifying elements blurred to ensure that no private information is compromised. The research did neither involve interventions with human subjects nor individualized human data. Thus, no approval from the Institutional Review Board was required by the author institutions.

Acknowledgments

This work was supported in part by the National Natural Science Foundation of China under Grant 62476247, 62073295 and 62072409, “Pioneer” and “Leading Goose” R&D Program of Zhejiang under Grant 2024C01214, in part by the Zhejiang Provincial Natural Science Foundation under Grant LR21F020003, in part by the Supcon Research Fund under Grant KYY-HX-20230833, and Lantai Research Fund under Grant KYY-HX-20240573, KYY-HX-20230365, KYY-HX-20250588.

References

- Burke, M.; Driscoll, A.; Lobell, D. B.; and Ermon, S. 2021. Using satellite imagery to understand and promote sustainable development. *Science*, 371(6535): eabe8628.
- Chen, L.; Li, J.; Dong, X.; Zhang, P.; He, C.; Wang, J.; Zhao, F.; and Lin, D. 2024a. Sharegpt4v: Improving large multi-modal models with better captions. In *European Conference on Computer Vision*, 370–387. Springer.
- Chen, M.; Li, Z.; Huang, W.; Gong, Y.; and Yin, Y. 2024b. Profiling urban streets: A semi-supervised prediction model based on street view imagery and spatial topology. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 319–328.
- Dong, H.; Li, J.; Wu, B.; Wang, J.; Zhang, Y.; and Guo, H. 2024. Benchmarking and improving detail image caption. *arXiv preprint arXiv:2405.19092*.
- Fan, L.; Krishnan, D.; Isola, P.; Katabi, D.; and Tian, Y. 2023a. Improving clip training with language rewrites. *Advances in Neural Information Processing Systems*, 36: 35544–35575.
- Fan, Z.; Zhang, F.; Loo, B. P.; and Ratti, C. 2023b. Urban visual intelligence: Uncovering hidden city profiles with street view images. *Proceedings of the National Academy of Sciences*, 120(27): e2220417120.
- Feng, J.; Wang, S.; Liu, T.; Xi, Y.; and Li, Y. 2025. UrbanLLaVA: A Multi-modal Large Language Model for Urban Intelligence with Spatial Reasoning and Understanding. *arXiv preprint arXiv:2506.23219*.
- Feng, J.; Zhang, J.; Yan, J.; Zhang, X.; Ouyang, T.; Liu, T.; Du, Y.; Guo, S.; and Li, Y. 2024. Citybench: Evaluating the capabilities of large language model as world model. *arXiv preprint arXiv:2406.13945*.
- Gao, P.; Han, J.; Zhang, R.; Lin, Z.; Geng, S.; Zhou, A.; Zhang, W.; Lu, P.; He, C.; Yue, X.; et al. 2023. Llama-adaptor v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.
- Hao, X.; Chen, W.; Yan, Y.; Zhong, S.; Wang, K.; Wen, Q.; and Liang, Y. 2025. Urbanvlp: Multi-granularity vision-language pretraining for urban socioeconomic indicator prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 28061–28069.
- He, K.; Fan, H.; Wu, Y.; Xie, S.; and Girshick, R. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9729–9738.
- He, K.; Zhang, X.; Ren, S.; and Sun, J. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 770–778.
- Huang, Y.; Zhang, F.; Gao, Y.; Tu, W.; Duarte, F.; Ratti, C.; Guo, D.; and Liu, Y. 2023. Comprehensive urban space representation with varying numbers of street-level images. *Computers, Environment and Urban Systems*, 106: 102043.
- Jean, N.; Wang, S.; Samar, A.; Azzari, G.; Lobell, D.; and Ermon, S. 2019. Tile2vec: Unsupervised representation learning for spatially distributed data. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, 3967–3974.
- Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.
- Lee, J.; Grosz, D.; Uzgent, B.; Zeng, S.; Burke, M.; Lobell, D.; and Ermon, S. 2021. Predicting livelihood indicators from community-generated street-level imagery. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 268–276.
- Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 12888–12900. PMLR.
- Li, J.; Selvaraju, R.; Gotmare, A.; Joty, S.; Xiong, C.; and Hoi, S. C. H. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in Neural Information Processing Systems*, 34: 9694–9705.
- Li, Q.; Ye, Z.; Feng, X.; Zhong, W.; Qin, L.; Chen, R.; Li, B.; Jiang, K.; Wang, Y.; Liu, T.; et al. 2025. CAI: Caption-Sensitive Attention Intervention for Mitigating Object Hallucination in Large Vision-Language Models. *arXiv preprint arXiv:2506.23590*.
- Li, Z.; Chai, Y.; Zhuo, T. Y.; Qu, L.; Haffari, G.; Li, F.; Ji, D.; and Tran, Q. H. 2023. Factual: A benchmark for faithful and consistent textual scene graph parsing. *arXiv preprint arXiv:2305.17497*.
- Li, Z.; Huang, W.; Zhao, K.; Yang, M.; Gong, Y.; and Chen, M. 2024. Urban Region Embedding via Multi-View Con-

- trastive Prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 8724–8732.
- Liang, Y.; Wen, H.; Xia, Y.; Jin, M.; Yang, B.; Salim, F.; Wen, Q.; Pan, S.; and Cong, G. 2025. Foundation Models for Spatio-Temporal Data Science: A Tutorial and Survey. *arXiv preprint arXiv:2503.13502*.
- Liu, F.; Chen, D.; Guan, Z.; Zhou, X.; Zhu, J.; Ye, Q.; Fu, L.; and Zhou, J. 2024. Remoteclip: A vision language foundation model for remote sensing. *IEEE Transactions on Geoscience and Remote Sensing*.
- Minderer, M.; Gritsenko, A.; and Hounsby, N. 2023. Scaling open-vocabulary object detection. *Advances in Neural Information Processing Systems*, 36: 72983–73007.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 8748–8763. PMLR.
- Sun, F.; Chang, Y.; Tanin, E.; Karunasekera, S.; and Qi, J. 2025. FlexiReg: Flexible Urban Region Representation Learning. *arXiv preprint arXiv:2503.09128*.
- Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024. Qwen2-vl: Enhancing vision-language model’s perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.
- Wu, S.; Yan, X.; Fan, X.; Pan, S.; Zhu, S.; Zheng, C.; Cheng, M.; and Wang, C. 2022. Multi-graph fusion networks for urban region embedding. *arXiv preprint arXiv:2201.09760*.
- Wu, W.; Zheng, K.; Ma, S.; Lu, F.; Guo, Y.; Zhang, Y.; Chen, W.; Guo, Q.; Shen, Y.; and Zha, Z.-J. 2024a. LoTLIP: Improving Language-Image Pre-training for Long Text Understanding. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 64996–65019.
- Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024b. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.
- Xi, Y.; Li, T.; Wang, H.; Li, Y.; Tarkoma, S.; and Hui, P. 2022. Beyond the first law of geography: Learning representations of satellite imagery by leveraging point-of-interests. In *Proceedings of the ACM Web Conference 2022*, 3308–3316.
- Xiao, C.; Zhou, J.; Xiao, Y.; Huang, J.; and Xiong, H. 2024. ReFound: Crafting a Foundation Model for Urban Region Understanding upon Language and Visual Foundations. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, 3527–3538.
- Yan, Y.; Wen, H.; Zhong, S.; Chen, W.; Chen, H.; Wen, Q.; Zimmermann, R.; and Liang, Y. 2024. Urbanclip: Learning text-enhanced urban region profiling with contrastive language-image pretraining from the web. In *Proceedings of the ACM on Web Conference 2024*, 4006–4017.
- Ye, Z.; Li, Q.; Feng, X.; Qin, L.; Huang, Y.; Li, B.; Jiang, K.; Xiang, Y.; Zhang, Z.; Lu, Y.; et al. 2025. CLAIM: Mitigating Multilingual Object Hallucination in Large Vision-Language Models with Cross-Lingual Attention Intervention. *arXiv preprint arXiv:2506.11073*.
- Yeh, C.; Perez, A.; Driscoll, A.; Azzari, G.; Tang, Z.; Lobell, D.; Ermon, S.; and Burke, M. 2020. Using publicly available satellite imagery and deep learning to understand economic well-being in Africa. *Nature Communications*, 11(1): 2583.
- Yong, X.; and Zhou, X. 2024. MuseCL: predicting urban socioeconomic indicators via multi-semantic contrastive learning. In *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence*, 7536–7544.
- Zhang, B.; Zhang, P.; Dong, X.; Zang, Y.; and Wang, J. 2024. Long-clip: Unlocking the long-text capability of clip. In *European Conference on Computer Vision*, 310–325. Springer.
- Zhang, L.; Long, C.; and Cong, G. 2022. Region embedding with intra and inter-view contrastive learning. *IEEE Transactions on Knowledge and Data Engineering*, 35(9): 9031–9036.
- Zheng, K.; Zhang, Y.; Wu, W.; Lu, F.; Ma, S.; Jin, X.; Chen, W.; and Shen, Y. 2024. Dreamlip: Language-image pre-training with long captions. In *European Conference on Computer Vision*, 73–90. Springer.
- Zhou, B.; Yang, H.; Chen, D.; Ye, J.; Bai, T.; Yu, J.; Zhang, S.; Lin, D.; He, C.; and Li, W. 2025. Urbanbench: A comprehensive benchmark for evaluating large multimodal models in multi-view urban scenarios. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, 10707–10715.
- Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.
- Zou, X.; Yan, Y.; Hao, X.; Hu, Y.; Wen, H.; Liu, E.; Zhang, J.; Li, Y.; Li, T.; Zheng, Y.; et al. 2025. Deep learning for cross-domain data fusion in urban computing: Taxonomy, advances, and outlook. *Information Fusion*, 113: 102606.